

# Strategies for two-stage sampling designs for estimating herd-level prevalence

Bruce Wagner<sup>a,\*</sup>, Mo D. Salman<sup>b</sup>

<sup>a</sup>*Centers for Epidemiology and Animal Health, U.S. Department of Agriculture, Animal and Plant Health Inspection Service, Veterinary Services, 2150 Centre Avenue, Fort Collins, CO 80526-8117, USA*

<sup>b</sup>*Animal Population Health Institute, College of Veterinary Medicine and Biomedical Sciences, Colorado State University, Fort Collins, CO 80523-1676, USA*

Received 29 July 2003; received in revised form 13 July 2004; accepted 13 July 2004

---

## Abstract

We propose a herd-level sample-size formula based on a common adjustment for prevalence estimates when diagnostic tests are imperfect. The formula depends on estimates of herd-level sensitivity and specificity. With Monte Carlo simulations, we explored the effects of different intracluster correlations on herd-level sensitivity and specificity. At low prevalence (e.g. 1% of animals infected), herd-level sensitivity increased with increasing intracluster correlation and many herds were classified as positive based only on false-positive test results. Herd-level sensitivity was less affected at higher prevalence (e.g. 20% of animals infected). A real-life example was developed for estimating ovine progressive pneumonia prevalence in sheep. The approach allows researchers to balance the number of herds and the total number of animals sampled by manipulating herd-level test characteristics (such as the number of animals sampled within a herd).

Published by Elsevier B.V.

**Keywords:** Herd-level prevalence; Two-stage sampling; Sensitivity; Specificity

---

\* Corresponding author. Tel.: +1 970 494 7256; fax: +1 970 494 7228.  
E-mail address: [Bruce.A.Wagner@aphis.usda.gov](mailto:Bruce.A.Wagner@aphis.usda.gov) (B. Wagner).

## 1. Introduction

Surveys of infection prevalence often are intended to substantiate claims of infection freedom at a regional or national level. A two-stage sampling approach has been proposed for such a purpose (Garner et al., 1997; Cameron and Baldock, 1998b; Audige and Beckett, 1999; Stärk et al., 2000). The first stage of sampling is the herd, and the second stage is within-herd sampling.

The use of two-stage sampling has evolved to meet surveillance objectives for two reasons. First, list frames of animals for randomized sample selection do not typically exist at a regional or national level—but list frames of herds can be constructed and maintained more readily. Secondly, the theory and application of within-herd sampling with imperfect diagnostic tests is well developed (Martin et al., 1992; Donald, 1993; Donald et al., 1994; Carpenter and Gardner, 1996; Jordan and McEwen, 1998; Cameron and Baldock, 1998a). The within-herd sampling research has guided the approach to sampling to classify the herd's disease or infection status.

The herd-level sensitivity (HSe) and specificity (HSp) depend on individual-animal test characteristics, sample size, within-herd prevalence, and the cutpoint number of reactors (Martin et al., 1992; Cameron and Baldock, 1998a). The HSe and HSp become test characteristics which can be applied at the herd level in a manner equivalent to animal-level sensitivity (Se) and specificity (Sp) at the within-herd level (Cameron and Baldock, 1998b). The HSe and HSp usually are based on detecting infection if it is present above a fixed level; the level is determined according to the epidemiology of the disease or specific national or international rules.

The determination of HSe and HSp, however, becomes difficult if a minimum within-herd prevalence cannot be assumed. When conducting a national survey to detect infection, it might be more appropriate to assume a distribution for within-herd prevalence of infection—especially as affected by clustering (Donald, 1993; Donald et al., 1994; Jordan and McEwen, 1998). Donald (1993) suggested modeling the distribution of within-herd prevalence as cluster-correlated binary responses to account for the nonindependence of infection status of animals within herds. He related the intracluster correlation coefficient to the beta-binomial distribution.

The U.S. National Animal Health Monitoring System (NAHMS) conducts national surveys of animal health and production (Wineland and Dargatz, 1998). One of the goals of NAHMS surveys is to estimate herd-level prevalences for various infections of interest in cattle, sheep, swine, horses, and poultry. Unfortunately, two-stage sampling methods for substantiating infection freedom are not directly applicable for estimating herd-level prevalence. Substantiating infection freedom relies on detection of infection at both the within-herd and herd levels. If the sample size needed for detecting infection at the herd level (or animal level) is used to design a study where a precise infection estimate is required, the resulting precision of the estimate might not be as precise as desired. The increased precision needed for an estimate as compared to sampling for detection typically will require a larger sample size.

Our objective was to develop two-stage sampling strategies for estimating herd-level prevalence given imperfect diagnostic tests, different herd sizes, and clustering of infection. We propose a sample-size formula for first-stage sampling and a Monte Carlo

simulation model was used to investigate the effect of infection clustering and various sampling strategies. Application of the approach is demonstrated with an example.

## 2. Materials and methods

### 2.1. Sample-size formula for herd-level estimation

Rogan and Gladen (1978) proposed a prevalence estimator that was adjusted for imperfect test characteristics. Although their estimator was intended for individual-test results, the estimator readily can be applied to herd-level estimation (Donald, 1993). The herd-level estimate of infection is:

$$\hat{\theta}_{RG} = \frac{\hat{\theta} + HSp - 1}{HSe + HSp - 1}$$

where  $\hat{\theta}_{RG}$  denotes the adjusted herd-level prevalence and  $\hat{\theta}$  denotes the apparent (i.e. unadjusted) herd-level prevalence.

Donald (1993) showed the variance of the estimator, assuming fixed HSe and HSp, is:

$$\text{var}(\hat{\theta}_{RG}) = \frac{\text{var}(\hat{\theta})}{(HSe + HSp - 1)^2}$$

The variance formula can be used to derive a sample-size equation (Appendix A). A sample-size estimate for the number of herds ( $n_h$ ) needed to achieve an acceptable error limit (error limit denotes one-half the maximum width for the confidence interval for  $\theta_{RG}$ ) with a specified confidence level ( $1 - \alpha$ ) is:

$$n_h = \frac{\theta(1 - \theta)}{(HSe + HSp - 1)^2} \left( \frac{Z_{\alpha/2}}{\text{error}} \right)^2$$

where  $\theta$  is the proportion of infected herds.

### 2.2. Simulation model characteristics

We constructed a model to perform a Monte Carlo simulation of two-stage sampling using a commercially available software package (@Risk, Palisade Corporation, Newfield, NY). Inputs into the model include herd size, animal-level sensitivity (Se) and specificity (Sp), animal-level prevalence, within-herd sample size, percent of herds that are negative, and the cutpoint needed to designate a herd as positive. Herd size can be input as a fixed size or it can follow a discrete distribution. Sensitivity and specificity can be input as point estimates or can be selected randomly from beta-pert distributions (minimum, maximum, and most likely values would be required) (Audige and Beckett, 1999). The within-herd sample size can be a fixed number, a fixed percent of the herd size, or based on the discrete distribution of herd size. The cutpoint is the number of test-positive results that are needed to designate a herd as “infected.” The default cutpoint is one. These input parameters are similar to those used by Audige and Beckett (1999) and Jordan and McEwen (1998).

When fixed parameters are input, results from the model follow those reported by [Cameron and Baldock \(1998b\)](#).

In each simulation iteration, the input parameters were used to create infected and noninfected herds. The model used a beta-binomial distribution to assign the number of infected animals to each herd. We chose the beta-binomial distribution because of its appropriateness for modeling prevalence within herds and at the herd level as well as its direct relationship to the intracluster correlation coefficient. The parameters for the beta-binomial distribution ( $a$  and  $b$ ) were determined by the specified animal-level prevalence ( $p$ ) and the intracluster correlation coefficient ( $\rho$ ) using the following formulas ([Bohning and Greiner, 1998](#)):

$$a = \frac{p}{\rho} - p$$

and

$$b = \frac{1-p}{\rho} + p - 1$$

An issue with using the continuous beta-binomial distribution is the possibility of “infected” herds not having any infected animals. When selecting the herd-level prevalence from the beta distribution, it is possible that a small prevalence level could be chosen. If the prevalence level is sufficiently low, the resulting selection from the binomial distribution used to determine the number of infected animals in the herd might be zero with a relatively high probability. One solution to the problem is to force at least a single infected animal into each “infected” herd, as was implemented by [Donald et al. \(1994\)](#). This option necessitates modeling a free and an infected population. A second option is to allow the simulation from the beta-binomial distribution to create both the infected and noninfected herds. Both of these options are available in our model and in the model developed by [Jordan and McEwen \(1998\)](#). We modeled infected and free populations when we investigated the effect of infection clustering and various sampling strategies since we could set the percent of herds that were not infected ( $\psi = 1 - \theta$ ). We chose to use the second option in the analysis of the real (observed field) data because we intended to estimate the intracluster correlation for use in the model and we did not have an estimate of the proportion of herds that were not infected.

After the number of infected animals in each herd was determined, the model used a hypergeometric probability distribution (to model sampling without replacement) to choose the number of infected animals in the sample from the herd randomly. The number of noninfected animals in the sample was the difference between the sample size and the number of infected animals in the sample. A binomial probability distribution (sampling with replacement) was used, as an approximation to the hypergeometric probability distribution, to estimate the number of infected animals in samples when the number of infected animals in larger herds was large.

The model used a binomial probability distribution to simulate the diagnostic testing of the sample of animals from the herd. Sensitivity and the number of infected animals defined the binomial probability distribution which then was used to determine randomly

the number of true test-positives and false test-negatives in each sample. Similarly, Sp was used to determine the number of true test-negatives and false test-positives.

For each combination of within-herd sample size, sensitivity, specificity, cutpoint, and intracluster correlation sampling and testing from 10,000 herds was simulated. The simulated test results for each herd were cross-classified into the traditional  $2 \times 2$  table based on the true herd status and the results from the simulated herd-level sampling and testing. This classification allowed for direct calculation of a point estimate of herd-level sensitivity and specificity based on simulated sampling of 10,000 herds. We chose to simulate 10,000 herds to allow for stable estimates of these characteristics.

### 2.3. Survey-model implementation

The model was used to investigate both hypothetical and real within-herd and herd-level sampling-design strategies. First, the effects of within-herd sample size, intracluster correlation coefficient, and cutpoint on HSe and HSp were assessed by simulation. For these simulations, herd-size was 200, Se = 90%, Sp = 90%, and proportion of the herds that was negative ( $\psi = 60\%$ ) was fixed. Six levels of intracluster correlation (0, 0.05, 0.1, 0.3, 0.5, and 0.8), two animal-level prevalences (1 and 20%), three within-herd sample sizes (10, 20, and 30) and three cutpoints for the number of test-positive animals in the sample needed to call a herd positive (1, 2, and 3) were used in the simulation. The impact of imperfect specificity was assessed by repeating the simulation with Sp = 100%. The values of the characteristics were chosen to allow for comparison of our model results to previously published work (Donald et al., 1994) and to represent a spectrum of realistic intracluster correlation coefficients.

Two-stage sampling then was simulated with the model for Se = 98%, Sp = 99%, and animal-level prevalence = 20%. The intracluster correlation coefficient ( $\rho = 0.1$ ) and the proportion of herds that were negative ( $\psi = 60\%$ ) was fixed. The relatively high sensitivity, specificity, and animal-level prevalence were selected to demonstrate the model under relatively benign conditions; issues of inability to attain desired HSe and HSp were left for other demonstrations. Two design scenarios were investigated to examine the changes in design when targeted herd-level test characteristics are altered. In the first simulated design scenario, within-herd sample sizes were determined separately for each herd-size category (10–30, 31–50, 51–100, and 101–200) with the constraint that the cutpoint be fixed at one. The within-herd sample size within each herd-size category was determined iteratively with the goal of HSe > 90% and the HSp > 80%. In the second simulated design scenario, the within-herd sample sizes were simulated again-but the cutpoint was allowed to vary. Under this design scenario, the sample size was determined to obtain approximately HSe = 95% and HSp = 90%. After the within-herd sample sizes were determined, each of the within-herd design scenarios was modeled for all herd-size categories to simulate two-stage sampling from a population of herds. The proportion of herds taken from each herd-size category was simulated in two ways for each of the within-herd sampling design scenarios. In the first herd-level sampling simulation, the proportion of herds from each herd-size category followed the proportion of herds in the population and ranged from 60% in the small-herd category to 5% in the large-herd category. In the second herd-level simulation, the proportion of herds sampled from each herd-size category was equivalent

(25% from each category) to represent a naïve design. The overall herd-level test characteristics were calculated after completion of the simulation of two-stage sampling from all herd sizes. These test characteristics represent the herd-level test characteristics that would be expected under the two-stage survey designs.

The herd-level test characteristics from the four combinations of within-herd and herd-level design scenarios were used as input into the sample-size equation to determine the number of herds needed to estimate herd-level prevalence with 95% confidence and 10% error limits. The confidence intervals were evaluated by simulating sampling of the required number of herds under each of the four design scenarios 1000 times to estimate accurately what proportion of the adjusted herd-level prevalence estimates fell within the confidence interval (e.g., if the number of herds required to be 95% confident with an error limit of 10% was 150, then 1000 simulations of sampling 150 herds were implemented). The simulation using 1000 iterations of design scenarios was implemented to obtain consistent results while maintaining a manageable output data set ( $1000 \times 150$ ). The simulation results were exported to a statistical software package (SAS, SAS Institute, Cary, NC) from which adjusted herd-level prevalence was calculated and descriptive statistics were computed.

As the last part of the two-stage simulation, the total number of animals needed to be sampled in each design scenario was calculated by multiplying the number of herds needed in each herd-size category by the within-herd sampling requirements for that size category. The number of herds sampled in each herd-size category depended on whether the herd-level sampling was proportional to the population or set to equivalent proportions.

#### 2.4. Survey-model application

We used the model to demonstrate the design of surveys to estimate herd-level prevalence for ovine progressive pneumonia (OPP) in sheep flocks. This example was chosen because data were available from a recent NAHMS study for estimating the input parameters of the model.

In the 2001 NAHMS sheep study, 21,525 sheep in 687 flocks were tested for OPP using a competitive-inhibition enzyme-linked immunosorbent assay (cELISA). The Se and Sp of the cELISA were estimated to be 98.6 and 96.9%, respectively (K. Marshall, personal communication). The proportion of herds in each herd-size category followed the herd allocation used by NAHMS. Only flocks with 20 or more animals were eligible for testing in the NAHMS study. Weighted animal-level prevalence was estimated from the data and adjusted using the Rogan–Gladden formula to obtain an animal-level prevalence estimate for the model. The intracluster correlation coefficient was estimated using hierarchical modeling software (MLwiN, version 1.1, Institute of Education, University of London, London, UK).

For comparison, two sampling designs were constructed for OPP in sheep. The first design restricted the cutpoint to a single positive for determining herd-level status. The sample size was chosen to keep HSp about 70%. The second design allowed flexible cutpoints for each herd-size category and attempted to keep HSe sensitivity about 80%. The HSp and HSe goals were selected to allow for comparison of sampling two different

within-herd sampling strategies—the former requires much fewer samples per herd than the latter.

The percent of herds that was negative was not used as an input parameter into the sheep-herd simulation. Instead, only  $\rho$  (calculated from the raw data) was used to determine the distribution of within-herd prevalence. All sampling designs were constructed to achieve an estimate of the herd-level prevalence with 95% confidence and an error limit of 10%.

### 3. Results

#### 3.1. Herd-level modeling results

At all sample sizes and values of  $\rho$ , when the animal-level test was imperfect, the effect of increasing cutpoint was to decrease HSe and increase HSp (Tables 1 and 2). Increasing sample size—regardless of cutpoint, animal-level prevalence, and  $\rho$ —increased HSe and decreased HSp. When the animal-level prevalence was relatively low ( $p = 1\%$ ), HSe increased with increasing  $\rho$ . When the within-herd prevalence was higher ( $p = 20\%$ ), the HSe varied less but appeared to decrease in the middle of the range of  $\rho$ . Assuming perfect animal-level test specificity substantially decreased HSe at both levels of animal-level prevalence. Additionally, trends across the levels of  $\rho$  were consistent with the results observed for imperfect specificity at the two animal-level prevalences examined.

Table 1

Effect of within-herd infection correlation ( $\rho$ ) with fixed herd size ( $N = 200$ ) and animal-level prevalence ( $p = 0.01$ ), and variable animal-level specificity, within-herd sample size ( $n$ ) and cut-off points ( $k$ ) on herd-level sensitivity and specificity

$n$	$k$	Herd-level sensitivity (%)						Herd-level specificity (%)
		$\rho = 0$	$\rho = 0.05$	$\rho = 0.1$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.8$	
Animal-test specificity = 0.9								
10	1	68.9	72.3	73.0	81.2	85.7	92.8	35.0
	2	29.8	34.7	38.1	47.6	57.1	69.7	73.7
	3	9.3	11.4	13.4	24.9	36.2	64.2	93.0
20	1	89.9	91.8	92.1	93.3	97.3	96.4	12.2
	2	66.1	70.6	72.3	81.2	85.5	85.8	39.4
	3	37.0	43.3	48.2	56.0	63.8	75.0	67.6
30	1	96.2	98.1	99.0	97.7	98.6	100.0	4.2
	2	85.1	88.8	91.0	89.5	92.8	94.3	18.4
	3	65.1	69.8	71.9	77.0	82.6	90.2	41.4
Animal-test specificity = 1.0								
10	1	10.3	19.5	26.5	34.1	46.9	65.3	100
20	1	19.7	31.2	40.1	45.7	59.9	76.9	100
30	1	27.0	39.8	51.4	55.6	64.7	78.9	100

Animal-test sensitivity = 0.9 and the proportion of negative herds = 0.6; 10,000 herds sampled.

Table 2

Effect of within-herd infection correlation ( $\rho$ ) with fixed herd size ( $N = 200$ ) and animal-level prevalence ( $p = 0.2$ ), and variable animal-level specificity, sample size ( $n$ ) and cut-off points ( $k$ ) on herd-level sensitivity and specificity

$n$	$k$	Herd-level sensitivity (%)						Herd-level specificity (%)
		$\rho = 0$	$\rho = 0.05$	$\rho = 0.1$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.8$	
Animal-test specificity = 0.9								
10	1	94.6	93.2	91.0	87.3	86.6	90.7	34.8
	2	77.6	74.4	71.5	66.8	68.9	77.0	73.8
	3	51.0	49.6	47.4	46.3	48.8	67.6	92.8
20	1	99.8	99.4	98.6	97.7	97.0	97.8	12.3
	2	98.2	96.0	93.7	88.1	87.4	91.5	39.2
	3	91.8	87.7	83.1	77.7	74.8	80.9	67.8
30	1	100.0	99.9	99.8	99.3	99.4	99.4	4.3
	2	99.9	99.2	98.3	95.8	95.4	96.1	18.4
	3	99.0	97.7	94.3	89.0	88.6	90.7	40.7
Animal-test specificity = 1.0								
10	1	85.5	79.8	75.7	65.7	64.7	75.7	100.0
20	1	98.0	93.9	88.8	75.8	74.9	79.2	100.0
30	1	99.8	97.5	94.1	82.6	78.5	83.7	100.0

Animal-test sensitivity = 0.9 and the proportion of negative herds = 0.6; 10,000 herds sampled.

### 3.2. Modeling two-stage sampling with a hypothetical population

In the sample design with a fixed cutpoint of one test-positive, the required within-herd sample size ranged from 12 to 20 animals per herd (Table 3). With these sample sizes, HSe and HSp were 91 and 81%, respectively, in all herd-size categories.

When the sample design allowed a cutpoint of one or more, the sample size was increased in all herd-size categories except the smallest herd size (where allowing a higher cutpoint made it impossible to attain an adequate HSe). In herds with  $>30$  animals, all animals were tested up to a maximum of 55. When the simulation was run with all herd-size categories at

Table 3

Variation in herd-level sensitivity and specificity for herd-size categories under two herd-level design strategies: with and without a fixed cutpoint of one

Herd size	Percent of herds	Cutpoint	Within-herd sampling	HSe (%)	HSp (%)
Fixed cutpoint					
10–30	60	1	All up to 12, then 13	93.3	87.7
31–50	20	1	15 up to 40, then 18	92.3	85.1
51–100	15	1	19	91.4	82.1
101–200	5	1	20	92.7	81.1
Variable cutpoint <sup>a</sup>					
10–30	60	1	All up to 12 then 13	93.3	87.7
31–50	20	2	Select all	95.8	93.6
51–100	15	1	All up to 55 then 55	97.1	89.7
101–200	5	1	55	95.8	89.4

Note: Se = 98%, Sp = 99%,  $p = 20\%$ ,  $\Psi = 60\%$ ,  $\rho = 0.1$ .

<sup>a</sup> Cutpoint = 1 for herd size  $\leq 30$ , then = 2 for herd sizes  $>30$ .

Table 4

Overall herd-level sensitivity (OHSe) and specificity (OHSp) for herd-size categories under two herd-level design strategies, with and without a fixed cutpoint of one, and two herd sampling strategies (proportion to the population and equivalent proportion in each herd-size category)

Herd sampling strategy	OHSe (%)	OHSp
Fixed cutpoint		
In proportion to population	93.4	85.8
25% of herds from each herd size category	93.0	85.8
Variable cutpoint <sup>a</sup>		
In proportion to population	95.1	90.7
25% of herds from each herd size category	94.3	90.1

Note: Se = 98%, Sp = 99%,  $p = 20\%$ ,  $\Psi = 60\%$ ,  $\rho = 0.1$ .

<sup>a</sup> Cutpoint = 1 up to herd size of 30 then = 2 for the rest of the herd sizes.

one time, the overall herd-level sensitivity and specificity (OHSe, OHSp) for the fixed cutpoint design was approximately 93 and 86%, respectively, regardless of whether the herds were sampled in proportion to the population or equal proportions were taken from each herd-size category (Table 4). The OHSe and OHSP for the variable-cutpoint design were close to 95 and 90%, respectively (again, regardless of the herd-sampling protocol).

The number of herds required to attain the desired confidence in estimation was between 127 and 152 (Table 5). The number of herds required for the fixed-cutpoint sampling was approximately 20 greater than was required by the variable-cutpoint sampling design. Adjusted herd-level prevalence estimates obtained from the model fell within 10% of the true prevalence at least 95% of the time—suggesting that the sample size was sufficient to create the desired confidence interval.

The within-herd sampling designs resulted in substantially varying number of animals that needed to be tested (Table 6). Although more herds were required for the fixed-cutpoint design, the total number of animals was less than required for variable-cutpoint designs. The variable-cutpoint design with equal proportions of samples taken from each herd-size category had the largest total sample size requirement because of the greater sample-size requirements in the three larger herd-size categories.

### 3.3. Modeling two-stage sampling: OPP example

Apparent (observed from the NAHMS study) within-flock prevalence of OPP in sheep flocks was highly variable (Fig. 1) which resulted in an estimated  $\rho$  of 0.48. Six flock-size categories were defined to provide flexibility for adjusting within-herd samples sizes. Almost 50% of the flocks had 100 or fewer adult ewes (Table 7). When a cutpoint was fixed, between 11 and 16 samples were required from each flock to keep HSp at approximately 70%. The HSp was lower in the largest flock size to avoid HSe from declining further. When the cutpoint was variable and the HSe objective was 80%, the HSp was between 82.1 and 47.9% and sample sizes increased to a maximum of 90 in the largest flocks. Overall herd-level test characteristics showed an improvement in all measures for the variable-cutpoint design compared to the fixed-cutpoint design.

The improved test characteristics of the variable-cutpoint design for OPP resulted in a smaller sample size of flocks than did the fixed-cutpoint design (Table 8). However, the

Table 5

Sample size needed to estimate herd-level prevalence (95% confident with an error limit of 10%), true prevalence, and simulated adjusted herd-level prevalence estimates under two herd-level design strategies, with and without a fixed cutpoint of one, and two herd sampling strategies (proportion to the population and equivalent proportion in each herd-size category)

Sample design and strategy	Number of herds required	Assumed true herd-level prevalence (%)	Adjusted prevalence <sup>a</sup> (%)		
			5th percentile	Mean prevalence	95th percentile
Cutpoint = 1 proportion to pop	150	38.1	28.6	36.6	44.8
Cutpoint = 1 (25% from each herd-size category)	152	37.1	30.7	39.2	47.5
Variable cutpoint, proportion to pop	127	38.8	29.1	37.3	45.3
Variable cutpoint (25% from each herd-size category)	131	37.1	30.4	38.7	42.5

<sup>a</sup> Distribution characteristics from 1000 simulations of sampling the specified number of herds.

Table 6

Estimated number of animals needed to estimate herd-level prevalence under two herd-level sampling strategies (fixed cutpoint of one and variable cutpoint)

Herd size	Mean number needed to be sampled per herd	Number of herds needed to be sampled	Number of animals
Cutpoint = 1, sample relative to proportion in population			
10–30	12.9	90	1161
31–50	16.5	30	495
51–100	19.0	22	399
101–200	20.0	8	160
Total		150	2215
Cutpoint = 1, sample equal proportion in each herd-size category			
10–30	12.9	38	490
31–50	16.5	38	627
51–100	19.0	38	722
101–200	20.0	38	760
Total		152	2599
Variable cutpoint, sample relative to proportion in population			
10–30	12.9	76	980
31–50	40.0	26	1040
51–100	54.7	19	1040
101–200	55.0	6	330
Total		127	3390
Variable cutpoint, sample equal proportion in each herd-size category			
10–30	12.9	33	426
31–50	40.0	33	1320
51–100	54.7	33	1807
101–200	55.0	32	1760
Total		131	5313

Within each strategy, herds were sampled either in proportion to how they occurred in the population or by a fixed 25% within each herd category.

increased number of animals that needed to be tested in each flock-size category resulted in more than double the number of animal tests than did the fixed-cutpoint design.

#### 4. Discussion

The design of two-stage sampling plans to estimate herd-level infection prevalence when the diagnostic test is imperfect is dependent on a number of variables which can be addressed in a model such as presented here. Often, a practical limitation of a modeling approach is the lack of empirical information upon which to base assumptions (Jordan and McEwen, 1998)—but the minimal requirements for this model are the distribution of herd sizes, reliable estimates of Se and Sp, and some information on the potential distribution of within-herd prevalence.

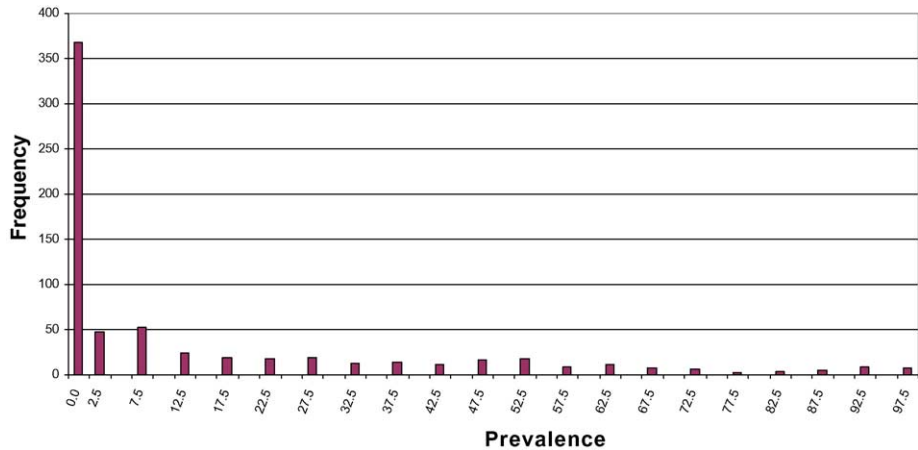


Fig. 1. Distribution of apparent within-flock prevalence of ovine progressive pneumonia in sheep flocks sampled in the 2001 National Animal Health Monitoring System’s sheep study.

The intracluster correlation coefficient along with the animal-level prevalence can be used to specify the distribution of within-herd prevalence in a population. Our results indicate that herd-level test characteristics can be influenced by the level of clustering in herds, especially at lower prevalences. McDermott and Schukken (1994) emphasized the

Table 7  
Herd-level sensitivity and specificity for herd-size categories under two herd-level design strategies (HSp = 70% and HSe = 80%), for testing for ovine progressive pneumonia in sheep flocks ( $\rho = 0.48$ )<sup>a</sup>

Herd size	Percent of herds	Within-herd sampling	HSe (%)	HSp (%)
Fixed cutpoint = 1, keeping herd-level specificity approximately 70%				
20–50	24	11	85.7	70.8
51–100	24	11	79.7	69.9
101–200	17	11 up to 150, then 12	75.3	69.3
201–500	15	12	73.1	69.7
501–2000	15	12	68.3	68.4
2001–5000	5	16	71.7	60.7
Sample of herds from population	–	–	76.3	70.0
Variable cutpoint <sup>b</sup> , keep herd-level sensitivity approximately 80%				
20–50	24	All up to 25, then 25	86.3	82.1
51–100	24	25 up to 80 then 30	81.5	79.9
101–200	17	30 up to 150, then 50	78.6	77.6
201–500	15	50 up to 250, then 60	78.9	72.9
501–2000	15	75	81.5	58.7
2001–5000	5	90	84.2	47.9
Sample of herds from population	–	–	81.4	77.6

Results are based on simulating 1000 herds in each herd-size category.  
<sup>a</sup> The presence of a noninfected population of herds was not assumed for this simulation—the noninfected herds were generated based on the intracluster correlation coefficient.  
<sup>b</sup> Cutpoint = 2 up to herd sizes of 150 then increases to 3.

Table 8

Sampling design for herd-level prevalence estimation of ovine progressive pneumonia for two herd-level sampling strategies (fixed cutpoint of one and flexible cutpoint)

Herd size	Mean number sampled per herd	Number of herds	Number of animals
Fixed cutpoint = 1, keeping herd-level specificity about 70%			
20–50	11.0	106	1166
51–100	11.0	104	1144
101–200	11.5	74	851
201–500	12.0	67	804
501–2000	12.0	67	804
2001–5000	16.0	21	336
Total		439	5105
Variable cutpoint, keep herd-level sensitivity about 80%			
20–50	24.5	66	1617
51–100	27.0	65	1755
101–200	40.0	46	1840
201–500	58.3	42	2450
501–2000	75.0	42	3150
2001–5000	90.0	13	1170
Total		274	11,982

need to consider cluster effects in epidemiological studies of animal populations and used the ANOVA method for estimating  $\rho$ . They also listed intracluster correlation coefficients from a number of studies on various species and diseases. We chose to estimate  $\rho$  using a hierarchical model. Estimates of  $\rho$  from actual test data will most likely be affected by test sensitivity and specificity. However, the estimate provides some basis for designing studies. Further research efforts are needed to develop adjusted estimates of  $\rho$ .

When no estimate of  $\rho$  is available, a qualitative assessment of  $\rho$  might be used. Ariwan and Frerichs (1996) suggested qualitative levels of “low,” “medium,” and “high” intracluster correlation which they associated with values of the design effect: 2, 4, and 7, respectively. The design effect is the variance of an estimated proportion obtained by a cluster sample divided by the variance for a simple random sample (Dargatz and Hill, 1996). Ariwan and Frerich’s (1996) software converts the design effect into an estimate of  $\rho$  using the following formula:

$$\rho = \frac{\text{deff} - 1}{\bar{n} - 1}$$

where deff: design effect and  $\bar{n}$ : average cluster (herd) size.

A common approach to determining whether a herd is “free” from infection is to design a sampling protocol to detect a minimum within-herd prevalence (Garner et al., 1997; Cameron and Baldock, 1998b; Cannon, 2001). For example, Garner et al. (1997) assumed that, in infected swine herds, at least 25% of the finisher pigs would have antibodies to the porcine reproductive-and-respiratory-syndrome virus. The minimal value can be considered to be the lower tail of the distribution of within-herd prevalence in a population that is not free from infection (such as would be expected in a herd-level infection-estimation problem). Assuming a minimal value will result in a conservative estimate of HSe. If the distribution of

the within-herd prevalence follows a “U” or “bathtub” shape (such as when  $\rho$  is relatively high), the HSe estimate can be too liberal if the design is based on the average within-herd prevalence. Thus, modeling the distribution of within-herd prevalence should provide a more-accurate assessment of overall herd-level test characteristics for sample size determination than does a detection level. Regardless of the approach, assumptions regarding the distribution of within-herd prevalence will have an impact on the design.

Both the hypothetical and the OPP examples for designing studies to estimate herd-level prevalence demonstrate that there are tradeoffs in choosing design strategies. The variable-cutpoint design substantially improved the overall herd-level specificity—while decreasing the number of herds that needed to be visited and increasing the number of animals for testing by  $\sim 30$ – $50\%$  (depending on the alternatives). For any study there are numerous possible approaches to the study design. The intent of the alternatives that we discussed in this paper were not to be comprehensive—but, rather, to demonstrate design flexibility and the need to consider underlying assumptions and tradeoffs during the design phase of a study to estimate herd-level prevalence. These tradeoffs give the designer or decision maker an option for evaluating the costs and objectives of proposed studies.

Herd-level prevalence estimation can be adjusted using the Rogan–Gladen method or through Bayesian techniques that are available (Enoe et al., 2000; Johnson et al., 2001). The potential for misclassification must be considered when risk factors are being evaluated. Greiner and Gardner (2000) discuss the misclassification at the individual-test level when the risk factor and the diagnostic classification both are not readily observable and surrogate measurements must be used. Unadjusted odds ratios systematically underestimate the true odds ratio when misclassification is nondifferential. When differential misclassification occurs, the odds ratio can be biased in either direction. However, differential misclassification can be a product of the design. For example, in the OPP variable-cutpoint design with the HSe close to 80%, the HSe was relatively stable by design—but the HSp ranged between 48% for large herds and 82% for smaller herds. Christensen and Gardner (2000) suggest that it is theoretically possible to adjust odds ratios in a risk-factor study if HSe and HSp are known. Alternatively, when the primary objective is to examine risk factors, some of the bias can be removed by adjusting the design to minimize differential misclassification.

The OHSeS in the hypothetical example were consistent whether the herds were sampled in proportion to the population or in equal proportions across the herd-size categories. This effect was due to the relatively homogenous HSeS and HSpS across the herd-size categories. If HSe and HSp varied dramatically across the herd-size categories, then altering the sampling proportions within herd-size categories could have an impact on the survey-level test characteristics. The prevalence estimates used for the sample-size calculations were slightly affected by the sampling protocol. This bias can be removed with a design-based analysis that assigns sampling weights to observations (Dargatz and Hill, 1996).

The model developed here has many similarities to models developed by Audige and Beckett (1999) and Jordan and McEwen (1998) but contains a combination of features not available in either of the other models. Our model allows for the use of a beta-binomial distribution for prevalence as does the Jordan model—but our model calculates the beta-distribution parameters based on intracluster correlation coefficients and animal-level prevalence. Additionally, several intracluster correlation coefficients can be evaluated

simultaneously. Our model differs from Audige's model in that our objective was to determine HSe and HSp under the assumption of infection clustering for use in estimating herd-level infection prevalence—but their objective was to assess freedom from infection. Another difference between our model and Audige's is our use of HSe and HSp to determine the number of herds to sample. Audige and Beckett (1999) developed herd-level test characteristics and then applied them to infected and noninfected herds in the second part of their model to build likelihood ratios for freedom versus an alternatively specified prevalence. Also, our model, unlike the other models, was developed to output results from the infection and testing simulation of individual herds so that herd-level testing strategies could be assessed readily.

Problems with using the beta-binomial distribution (of which the parameters are estimated  $\rho$  and animal-level prevalence) to model clustered populations have been noted by others. The probability of zero prevalence in the continuous distribution is zero—so that noninfected herds are not probabilistically plausible. Donald (1993) and Donald et al. (1994) recommended dividing the population under study into infected and noninfected populations. Herd-level specificity applies only to noninfected herds and HSe applies to the infected herds. A minimum of a single positive animal is forced into infected herds if the modeled prevalence is too low to assign one based on binomial probabilities. Audige and Beckett (1999) used a similar strategy while Jordan and McEwen (1998) offered an option for forcing the minimum of a single positive into infected herds. Our model has the flexibility to adopt either choice. We used the approach of dividing the population into infected and noninfected for the hypothetical modeling and chose not to force a single positive animal into “positive” herds. The modeling of OPP did not assume the presence of an infected and noninfected population.

## Acknowledgements

The study was supported by the USDA: Cooperative State Research, Education, and Extension Service through the Colorado State University – Center for Economically Important Infectious Animal Diseases and by the Research Council of the College of Veterinary Medicine and Biomedical Sciences, Colorado State University. From a thesis submitted to the Academic Faculty of Colorado State University in partial fulfillment for the degree of Doctor of Philosophy.

## Appendix A

The variance of the Rogan–Gladen estimator can be rewritten as follows (Donald, 1993):

$$\text{var}(\hat{\theta}_{\text{RG}}) = \frac{\text{var}(\hat{\theta})}{(\text{HSe} + \text{HSp} - 1)^2} = \frac{(\hat{\theta}(1 - \hat{\theta}))/n_h}{(\text{HSe} + \text{HSp} - 1)^2} = \frac{\hat{\theta}(1 - \hat{\theta})}{n_h(\text{HSe} + \text{HSp} - 1)^2}$$

where  $\hat{\theta}$ : unadjusted estimate of herd-level prevalence, HSe: herd-level sensitivity, HSp: herd-level specificity and  $n_h$ : herd sample size.

The variance estimate can be used to estimate the limits on error associated with an estimate. The usual formula for the limit on error for a sample proportion,  $\theta$ , is:

$$\text{error} \leq Z_c \sqrt{\frac{\theta(1-\theta)}{n}}$$

where  $Z_c$  is the value from standard normal distribution corresponding to confidence level  $c$ . Thus, the limit in error of the Rogan–Gladen estimator provides the following formula:

$$\text{error} \leq Z_{\alpha/2} \sqrt{\frac{\theta(1-\theta)}{n_h(\text{HSe} + \text{HSp} - 1)^2}}$$

Solving for  $n_h$  results in the following formula for estimating herd-level sample size:

$$n_h = \frac{\theta(1-\theta)}{(\text{HSe} + \text{HSp} - 1)^2} \left( \frac{Z_{\alpha/2}}{\text{error}} \right)^2$$

## References

- Ariwan, I., Frerichs R.R., 1996. User's Manual: CSurvey, Version 1.5. University of Indonesia, Indonesia and University of California, Los Angeles.
- Audige, L., Beckett, S., 1999. A quantitative assessment of the validity of animal-health surveys using stochastic modeling. *Prev. Vet. Med.* 38, 259–276.
- Bohning, D., Greiner, M., 1998. Prevalence estimation under heterogeneity in the example of bovine trypanosomosis in Uganda. *Prev. Vet. Med.* 36, 11–23.
- Cameron, A.R., Baldock, F.C., 1998a. A new probability formula for surveys to substantiate freedom from disease. *Prev. Vet. Med.* 34, 1–17.
- Cameron, A.R., Baldock, F.C., 1998b. Two-stage sampling in surveys to substantiate freedom from disease. *Prev. Vet. Med.* 34, 19–30.
- Cannon, R.M., 2001. Sense and sensitivity—designing surveys based on an imperfect test. *Prev. Vet. Med.* 49, 141–163.
- Carpenter, T.E., Gardner, I.A., 1996. Simulation modeling to determine herd-level predictive values and sensitivity based on individual-animal test sensitivity and specificity and sample size. *Prev. Vet. Med.* 27, 57–66.
- Christensen, J., Gardner, I.A., 2000. Herd-level interpretation of test results for epidemiologic studies of animal diseases. *Prev. Vet. Med.* 45, 83–106.
- Dargatz, D.A., Hill, G.W., 1996. Analysis of survey data. *Prev. Vet. Med.* 28, 225–237.
- Donald, A., 1993. Prevalence estimation using diagnostic tests when there are multiple, correlated disease states in the same animal or farm. *Prev. Vet. Med.* 15, 125–145.
- Donald, A.W., Gardner, I.A., Wiggins, A.D., 1994. Cut-off points for aggregate herd testing in the presence of disease clustering and correlation of test errors. *Prev. Vet. Med.* 19, 167–187.
- Enoe, C., Georgiadis, M.P., Johnson, W.O., 2000. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Prev. Vet. Med.* 45, 61–81.
- Garner, M.G., Gleeson, L.J., Holyoake, P.K., Cannon, R.M., Doughty, W.J., 1997. A national serological survey to verify Australia's freedom from porcine reproductive and respiratory syndrome. *Aust. Vet. J.* 75, 596–600.
- Greiner, M., Gardner, I.A., 2000. Application of diagnostic tests in veterinary epidemiologic studies. *Prev. Vet. Med.* 45 (1–2), 43–59.
- Johnson, W.O., Gastwirth, J.L., Pearson, L.M., 2001. Screening without a “gold standard”: the Hui-Walter paradigm revisited. *Am. J. Epidemiol.* 153, 921–924.
- Jordan, D., McEwen, S.A., 1998. Herd-level test performance based on uncertain estimates of individual test performance, individual true prevalence and herd true prevalence. *Prev. Vet. Med.* 36, 187–209.

- Martin, S.W., Shoukri, M., Thorburn, M.A., 1992. Evaluating the health status of herds based on tests applied to individuals. *Prev. Vet. Med.* 14, 33–43.
- McDermott, J.J., Schukken, Y.H., 1994. A review of methods used to adjust for cluster effects in explanatory epidemiological studies of animal populations. *Prev. Vet. Med.* 18, 155–173.
- Rogan, W.J., Gladen, B., 1978. Estimating prevalence from the results of a screening test. *Am. J. Epidemiol.* 107, 71–76.
- Stärk, K.D., Mortensen, S., Olsen, A.M., Barfod, K., Botner, A., Lavritsen, D.T., Strandbygaard, B., 2000. Designing serological surveillance programmes to document freedom from disease with special reference to exotic viral diseases of pigs in Denmark. *Rev. Sci. Tech.* 19, 715–724.
- Wineland, N.E., Dargatz, D.A., 1998. The National Animal Health Monitoring System. A source of on-farm information. *Vet. Clin. N. Am. Food Anim. Pract.* 14, 127–139.